CS 7545: Machine Learning Theory **Fall 2018**

Lecture 18: Contextual Bandits

Lecturer: Jacob Abernethy *Scribes: Vishvak S Murahari*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

18.1 Introduction

In the stochastic bandit setting, the algorithm picks an arm based on previous rewards. The contextual bandit is a more specific version of this setting, where the algorithm receives a context c_t on every round. The setting is formally defined below,

Setting

- In rounds $t = 1, \ldots, k$
  - Nature reveals context c_t.
  - Algorithm plays action a_t.
  - Nature reveals reward X_t, the reward for playing action a_t

Some scenarios where this framework might be useful:

Movie Recommendations In this scenario, the context could be features and the actions could be the set of all movies.

Path Recommendation In this scenario, the context could be a starting and ending point and the actions could be the set of all paths.

18.2 Notion of Regret

$$\text{Regret}_T = \mathbb{E}[\sum_{c \in C} \max_{k \in K} \sum_{t \in [T]; c_t = c} (x_{tk} - x_t)],$$

We recall that the regret for EXP3 algorithm for T rounds and K arms is:

$$\text{Regret}_T^{EXP} = \sqrt{TKlogK}$$

Therefore, if we assume a fixed context we can use the EXP3 bound to upper bound the regret,

$$\text{Regret}_{T.c} = \mathbb{E}[\max_{k \in K} \sum_{t \in [T]; c_t = c} (x_{tk} - x_t)],$$

$$\leq \sqrt{KlogK \sum_{t \in T} 1[c_t = c]} \leq \sqrt{KlogKT}$$

If we assume that all the contexts are equiprobable, we get the following bound:

$$\text{Regret}_{T.c} \leq \sqrt{\frac{TKlogK}{|C|}}$$

$$\text{Regret}_T = \sum_{c \in C} \text{Regret}_{T,c} \leq \sqrt{TK|C|logK}$$

18.3 Algorithms

We will construct an algorithm by using expert predictions. We will define M experts, $\phi_1, \phi_2 \dots \phi_M$, where $\phi_i : C- > \Delta(K)$. Therefore,

$$\text{Regret}_T = \mathbb{E}[(\max_{m \in M} \sum_{t=1}^{T} E^t_m.x_t) - \sum_{t=1}^{T} X_t],$$

Where,

E^t: Prediction of all experts

$E^t_{m,k}$: Probability that expert m suggests to use action k on round t

x^t: Reward vector of size K

X^t: Scalar reward received by playing action a_t

We will now present a new algorithm, EXP4 (Exponential Weighting for Exploration and Exploitation with Experts) and we will show that this algorithm achieves a tighter upper bound on the Regret.

Algorithm 1: EXP4

Input : T,K,M,N
$Q_1 = (\frac{1}{M}, \frac{1}{M}, \frac{1}{M} \dots, \frac{1}{M})$
for $t = 1, \dots, T$ **do**
  Receive advice E^t
  $p^t = Q_t E^t$
  $A^t \sim p^t$
  Play A^t
  $\hat{x_{ti}} = 1 - \frac{1[A_t = i](1 - X_t)}{p_{ti}}$
  $\widetilde{x_t} = E^t.\hat{x_t}$
  $Q_{t+1,i} = \frac{exp(\eta.\widetilde{X_{ti}}).Q_{ti}}{\sum_{j=1}^{M} exp(\eta \widetilde{X_{tj}}).Q_{tj}}$

18.4 Analysis

We will prove that the Regret for the EXP4 algorithm has the following bound:

$$\text{Regret}_T \leq \sqrt{2TKlog(M)},$$

We assume the following result for any $m^* \in M$:

$$\sum_{t=1}^{T} x_{\tilde{tm^*}} - \sum_{t=1}^{T} \sum_{m=1}^{M} Q_{tm} \widetilde{x_{tm}} \leq \frac{log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{m=1}^{M} Q_{tm} (1 - x_{\tilde{tm}})^2$$

Let,

$$m^* = \text{argmax}_{m \in M} \sum_{t=1}^{T} E^t_m.x_t$$

$$\mathbb{E}[\hat{x_t}] = x_t$$

$$\mathbb{E}[\widetilde{x_t}] = \mathbb{E}[E^t\hat{x_t}] = E^t\mathbb{E}[\hat{x_t}] = E^t.x_t$$

$$\sum_{t=1}^{T} x_{\tilde{tm^*}} - \sum_{t=1}^{T}\sum_{m=1}^{M} Q_{tm}\widetilde{x_{tm}} \leq \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{m=1}^{M} Q_{tm}(1 - x_{\tilde{tm}})^2$$

Therefore,

$$R_T \leq \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}[Q_{tm}(1 - x_{\tilde{tm}})^2]$$

We also assume the following result:

$$\mathbb{E}[(1 - x_{\tilde{tm}})^2] \leq \sum_{i=1}^{K}\frac{E_{mi}^t}{p_{ti}}$$

Therefore,

$$R_T \leq \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\sum_{m=1}^{M} Q_{tm}\sum_{i=1}^{K}\frac{E_{mi}^t}{p_{ti}}]$$

$$= \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\sum_{i=1}^{K}\sum_{m=1}^{M} Q_{tm}\frac{E_{mi}^t}{p_{ti}}]$$

$$= \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\sum_{i=1}^{K}\frac{1}{p_{ti}}\sum_{m=1}^{M} Q_{tm}E_{mi}^t]$$

$$= \frac{log(M)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\sum_{i=1}^{K}\frac{1}{p_{ti}}p_{ti}]$$

Therefore,

$$R_T \leq \frac{log(M)}{\eta} + \frac{\eta}{2}KT$$

Setting $\eta = \sqrt{\frac{2logM}{TK}}$, we can get the following bound

$$R_T \leq \sqrt{2TKlogM}$$